SUPPLEMENTARY MATERIAL

SUPPLEMENTARY MATERIALS AND METHODS

Gut microbial community DNA preparation

Fecal samples from humanized mice were stored at -80˚C before processing. DNA was extracted by bead-beating followed by phenol-chloroform extraction as described previously (*S1*). Luminal samples were collected from small intestinal and colonic segments by gently flushing each segment with 500 μL of Buffer A [200 mM Tris (pH 8.0), 200 mM NaCl, 20 mM EDTA], within 30 min after the animal had been sacrificed. The perfusate was then subjected to the same bead-beating, phenol-chloroform extraction procedure as used for the fecal samples. Small intestinal and colonic segments were subsequently opened along their cephalocaudal axis using a sterile scissors, and mucosa-adherent communities harvested by gently swabbing a 1 μL capacity sterile plastic loop along the length of each segment in a single motion. The contents of the loop were placed in 500 μL of Buffer A and DNA was prepared. Stomach and cecal contents were collected by manual extrusion, and the tissue was rinsed thoroughly in PBS before swabbing with the loop.

Sequencing of 16S rRNA gene amplicons

The V2 region (primers 8F-338R) was targeted for amplification and multiplex FLX pyrosequencing (*S1*). V2 16S rRNA gene sequencing data were pre-processed to remove sequences with low quality scores, sequences with ambiguous characters, and sequences outside the length bounds (200-300 nucleotides). They were then binned according to their sample-specific error-correcting barcode that was incorporated into the reverse primer. Similar sequences were identified using cd-hit (*S2*), with the following parameters: minimum coverage 99% and minimum pairwise identity 97% [see (*S1, S3, S4*) for more information on phylotype binning (OTU picking), UniFrac-based clustering, tree building, and rarefaction].

Full-length bacterial 16S rRNA gene sequence-based surveys were conducted through PCR amplification, cloning, sequencing, and analysis (including chimera-checking) as previously described (*S1*).

Isolation of *C. innocuum* strain SB23

After over two months on the Western diet, a single humanized gnotobiotic mouse was sacrificed and its cecal contents obtained under strictly anaerobic conditions. Contents were diluted 10^{-5} and 10^{-7} in rich medium [MBC plus 1% glucose; (*S5*)], plated on brain-heart infusion (BHI; Beckton Dickinson) agar supplemented with 10% horse blood (Colorado Serum Co.), and grown anaerobically at 37°C. Individual colonies were picked, screened by RFLP plus 16S rRNA gene sequencing, and saved as glycerol stocks.

Pyrosequencing of total community DNA and the *C. innocuum* SB23 genome

Shotgun sequencing runs were performed on the 454 FLX pyrosequencer from total fecal community DNA. Eleven microbiomes were analyzed in a single run using the default 454 FLX multiplex identifiers (MIDs), while the *C. innocuum* SB23 genome was sequenced alone in a full FLX run and a partial Titanium sequencing run (1/3 of a picotiter plate). Sequencing reads with degenerate bases ('Ns') were removed along with all replicate sequences [http://microbiomes.msu.edu/replicates using the following parameters: 0.9% identity, length difference requirement=0, and 3 bases checked (*S6*)].

The SB23 genome was assembled using Newbler v2.0.00.22 with the parameters '-g -consed' and all large contigs were used for the subsequent analyses.

C. innocuum strain SB23 genome annotation and metabolic reconstruction

The genome was annotated on-line using the RAST annotation pipeline (*S7*). Predicted proteins were also annotated based on the CAZy database [CAZy; www.cazy.org (*S8*)], with manual inspection. Predicted phage were identified using 'ProphageFinder' (e-value<0.001, 10 hits per prophage, hit spacing=5500, tRNAScan=on) (*S9*) and putative CRISPR regions were identified with 'CRISPRFinder' (*S10*). Percent identity plots were generated using the MUMmer3.20 package (*S11*).

Database searches and *in silico* microbiome metabolic reconstructions

The distributions of taxa, genes, orthologs, metabolic pathways, and higher-level gene categories in sequenced microbiomes were tallied based on the corresponding annotation of the best-BLAST-hit sequence found in each reference database (BLASTX e-value<10^{-5}, %identity>50, and score>50). For KEGG analysis, the closest matching gene with an annotation was used, since many genes in the database remain unannotated. When multiple annotated genes with an identical e-value were encountered after a BLAST query, we tallied all KEGG orthologous groups (KOs) assigned to those genes (NCBI BLASTX parameters: -e 10^{-5} -m 9 -b 100). Custom Perl scripts were used for all KEGG (v44) analyses (*S12*).

All microbiomes were aligned to a custom database of 122 reference human gut genomes in a manner that was similar to that used for transcriptome data (see section below regarding 'meta-transcriptomics'). Briefly, sequences were mapped with SSAHA2 (*S13*) to this database (**table S9**; SSAHA2 parameters: -best 1 -454) and reads were assigned based on best-match (SSAHA2 score≥100). Each gene's abundance value was normalized by z-score, a correlation distance matrix was constructed, clustered with UPGMA, and displayed in a dendrogram format (Matlab version 7.7.0).

Meta-transcriptomics (RNA-Seq)

Methods for microbial mRNA sequencing were recently developed and validated as part of a separate study using sequenced human gut bacterial species that were cultured *in vitro* and were harvested from colonized gnotobiotic mice.

Cecal samples from the mice described in the present study were stored at -80˚C before processing. An aliquot (~100mg) of each sample was suspended, while frozen, in a solution containing 500 µL of extraction buffer [200 mM NaCl, 20 mM EDTA], 210 µL of 20% SDS, 500 µL of a mixture of phenol:chloroform:isoamyl alcohol (pH 4.5, 125:24:1, Ambion 9720), and 250 µL of a slurry of acid-washed glass beads (Sigma-Aldrich). Microbial cells were subsequently lysed by mechanical disruption with a bead beater (BioSpec Products) set on high for 5 min at room temperature, followed by extraction with phenol:chloroform:isoamyl alcohol (pH 4.5, 125:24:1), and precipitation with isopropanol. RNA was purified using MEGAClear columns (Ambion). Total RNA was depleted of rRNA using MICROBExpress (Ambion) in addition to custom biotinylated oligonucleotides, directed against conserved regions of sequenced human gut bacterial rRNA genes, for streptavidin-based pulldown. cDNA was then synthesized using SuperScript II and random hexamers (Invitrogen), followed by second strand synthesis with RNAseH and *E.coli* DNA polymerase (New England Biolabs). Samples were prepared for sequencing with an Illumina GAII instrument after sonication in a BioRuptor XL sonicator (Diagenode). Reads produced from the each run were trimmed at their beginning and ends to remove bases with a quality score less than 20, and adapter sequences and sequencing reads with a resulting total length less than 20 nucleotides were eliminated from further analysis.

All trimmed reads were mapped with SSAHA2 (*S13*) to the *C. innocuum* SB23 genome and to the database of 122 reference human gut microbial genomes (SSAHA2 parameters: -best 1 -score 20 -solexa; **table S9**), sequenced largely through the sponsorship of the Human Gut Microbiome Initiative (HGMI, http://genome.wustl.edu/). Adaptor, rRNA, and tRNA sequences were removed prior to further analysis.

The number of transcripts assigned to each gene in the *C. innocuum* SB23 genome was tallied and normalized to reads per kilobase per million mapped reads (RPKM) (*S14*). SSAHA2 outputs against the 122 reference human gut genome database were grouped into gene clusters prior to counting [clusters were defined using the program cd-hit [parameter -c 0.8 (*S2*)] and the protein sequences encoded by predicted genes present in this database. Gene cluster counts were normalized based on 'counts per million'. Reads with significant homology (BLASTN e-value<10^{-30}) to annotated non-coding transcripts from the 122 gut genomes were excluded from subsequent analyses. To account for genes and gene clusters that were not detected due to limited sequencing depth, a 'pseudocount' of 1 was added to all samples.

Each gene's expression values were normalized by z-score, a correlation distance matrix was constructed, clustered with UPGMA, and displayed in dendrogram format (Matlab version 7.7.0). Sets of significantly enriched or depleted genes were identified using Cyber-T (*S15*).

qRT-PCR analysis

Total community RNA was extracted and cDNA was synthesized from samples taken from humanized mice fed the LF/PP or the Western diet (**fig. 1B**; N=3-5 samples/group; see the previous section for details). Primers for selected SB23 genes were designed using Primer3 (*S16*) and checked for specificity using (i) SSAHA2 comparisons against the 122-genome database, (ii) PCR against SB23 and community DNA, and (iii) qRT-PCR dissociation curves. qRT-PCR assays were run using ABsolute^{TM} QPCR SYBR$^{®}$ Green ROX Mix (Thermo Scientific) on a Mx3000P QPCR System instrument (Stratagene, La Jolla, CA). Fold-changes were calculated relative to the 16S rRNA gene using the $2^{-\Delta\Delta Ct}$ method: primers [1111F (5' - AACCCTTGTCGCATGTTACC - 3') and 1269R (5' – TCACTGTGTCGCTGCTCTTT - 3')].

Statistical analyses

Xipe (*S17*) (version 2.4) was used for bootstrap analyses of pathway enrichment and depletion, using the parameters sample size = 10,000 and confidence level = 0.95. Student's t-tests were used to identify statistically significant differences between two groups (Excel version 11.0, Microsoft). The Bonferroni correction was used to correct for multiple hypotheses.

SUPPLEMENTARY REFERENCES

S1. P. J. Turnbaugh *et al.*, A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2009).

S2. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

S3. R. E. Ley *et al.*, Evolution of mammals and their gut microbes. *Science* **320**, 1647-1651 (2008).

S4. N. Fierer, M. Hamady, C. L. Lauber, R. Knight, The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci U S A* **105**, 17994-17999 (2008).

S5. B. S. Samuel, J. I. Gordon, A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism. *Proc Natl Acad Sci U S A* **103**, 10011-10016 (2006).

S6. V. Gomez-Alvarez, T. K. Teal, T. M. Schmidt, Systematic artifacts in metagenomes from complex microbial communities. *Isme J*, (2009).

S7. R. K. Aziz *et al.*, The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

S8. B. L. Cantarel *et al.*, The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **37**, D233-238 (2009).

S9. M. Bose, R. D. Barber, Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol* **6**, 223-227 (2006).

S10. I. Grissa, G. Vergnaud, C. Pourcel, CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**, W52-57 (2007).

S11. S. Kurtz *et al.*, Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).

S12. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277-280 (2004).

S13. Z. Ning, A. J. Cox, J. C. Mullikin, SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-1729 (2001).

S14. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).

S15. P. Baldi, A. D. Long, A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17**, 509-519 (2001).

S16. S. Rozen, H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).

S17. B. Rodriguez-Brito, F. Rohwer, R. A. Edwards, An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**, 162 (2006).

S18. J. R. Cole *et al.*, The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* **33**, D294-296 (2005).

S19. P. J. Turnbaugh, F. Backhed, L. Fulton, J. I. Gordon, Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe* **3**, 213-223 (2008).

S20. D. Posada, K. A. Crandall, MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-818 (1998).

S21. D. L. Swofford. Phylogenetic Analysis Using Parsimony (*and Other Methods). (Sinauer Associates, Sunderland, Massachusetts, 2003).

S22. R. Barrangou *et al.*, CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712 (2007).

SUPPLEMENTARY FIGURES

Fig. S1. Phylogeny of the Erysipelotrichi. 16S rRNA gene sequences for Erysipelotrichi strains isolated from the human gut were identified in the RDP database (*18*). Near full-length representative sequences from the Western diet-associated Erysipelotrichi bloom in Western diet-fed mice with a native mouse gut microbiota (*S19*) and in gnotobiotic mice with a transplanted human gut microbiota (**table S2**) were selected from an ARB neighbor-joining tree. Likelihood parameters were determined using Modeltest (*S20*) and a maximum-likelihood tree was generated using PAUP (*S21*). Bootstrap values represent nodes found in >70 of 100 repetitions. *Eubacterium rectale,* a member of the Clostridia, was used to root the tree.

Turnbaugh *et al.*, Supplementary Figure 1

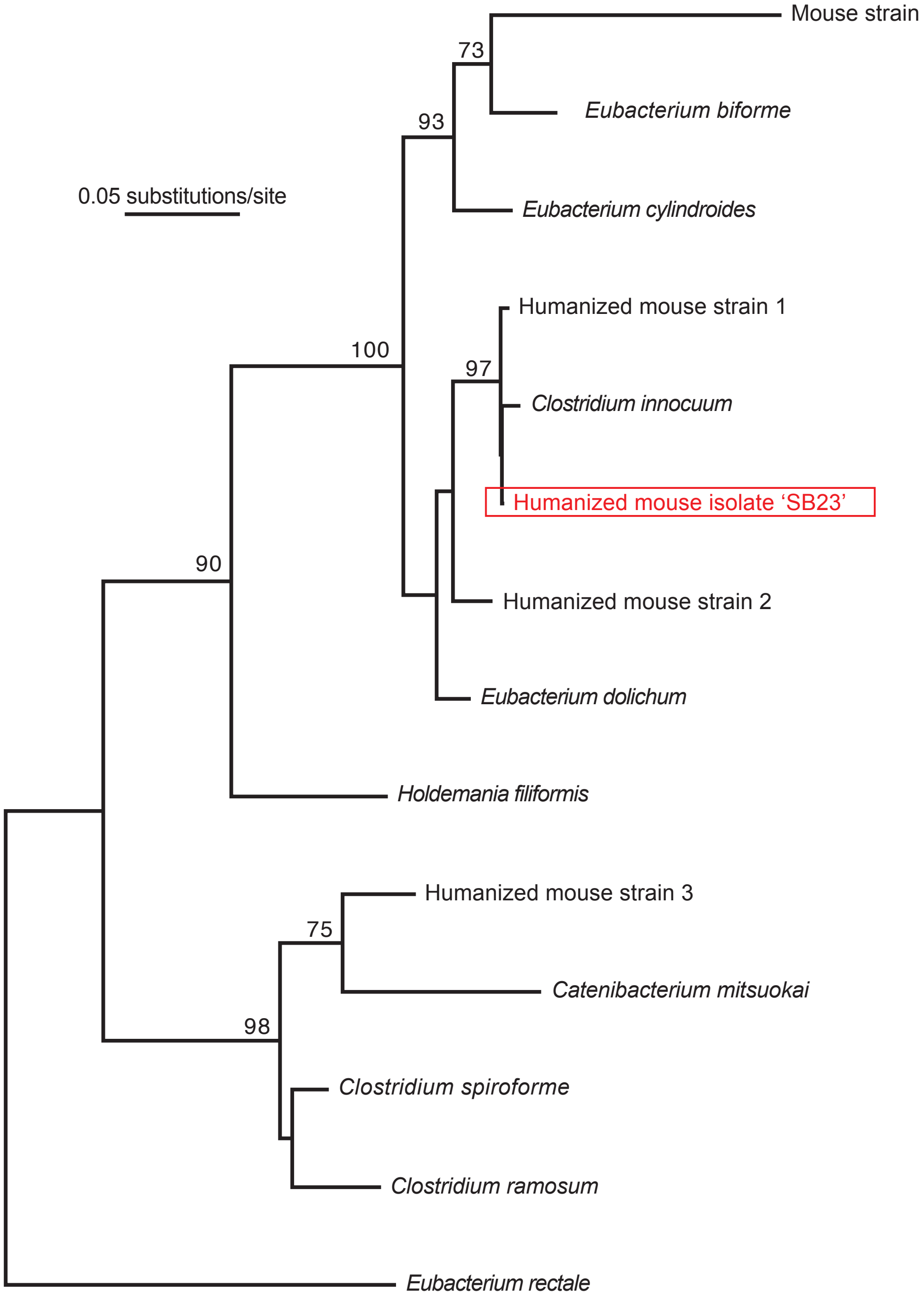


Fig. S2. Rarefaction analysis of humanized mice. Rarefaction curves are based on V2 16S rRNA gene sequences from (**A**) the freshly voided human fecal sample, and two generations of humanized mice on the LF/PP or Western diets, (**B**) the frozen human fecal subsample, and fecal samples obtained from LF/PP diet-fed mice 0.5-49 days after colonization with the previously frozen subsample, and (**C**) material taken from the stomach, small intestine, cecum, colon, and feces of humanized mice (species-level phylotypes defined by ≥97% identity).

Turnbaugh *et al.*, Supplementary Figure 2

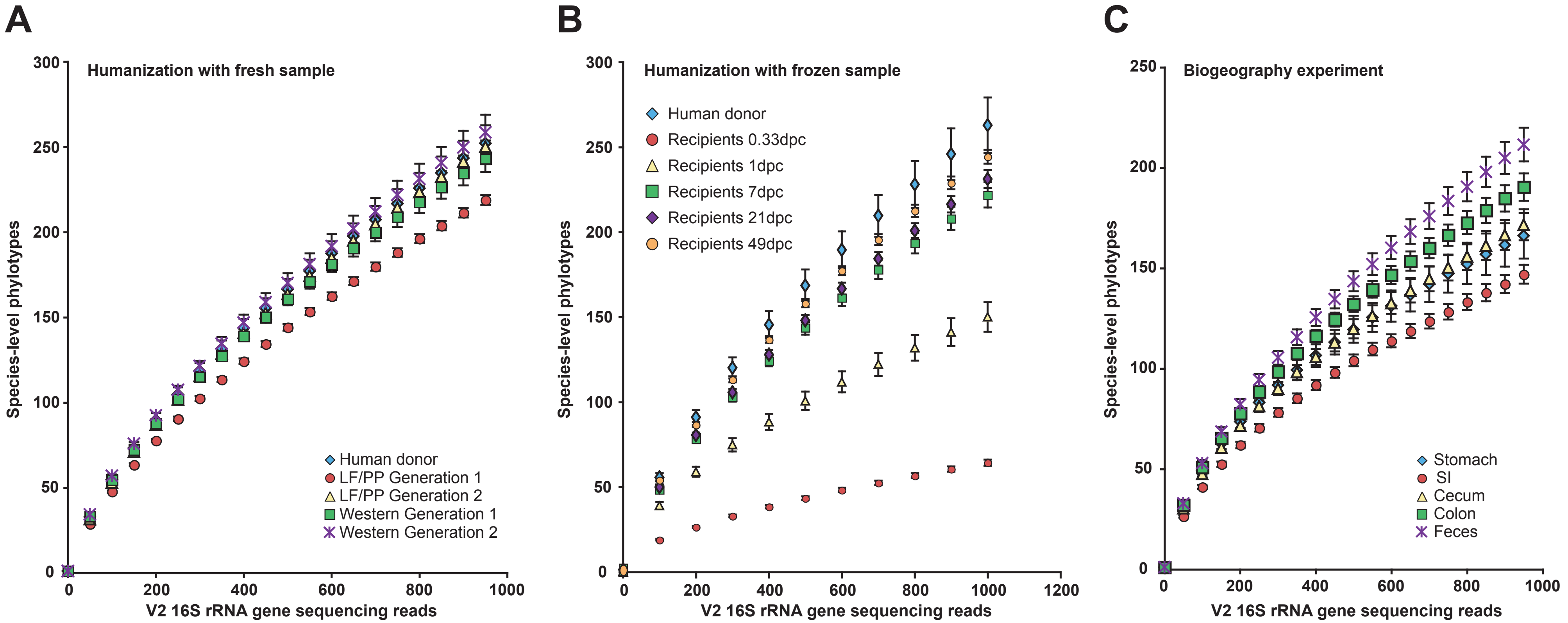


Fig. S3. Unweighted UniFrac-based clustering of V2 16S rRNA gene surveys. Samples were taken from the human donor (green), first-generation humanized mice fed LF/PP (red) or Western (blue) diets, second generation human microbiota transplant recipients consuming the LF/PP (light blue) or Western (purple) diets, and mice humanized with a frozen sample fed LF/PP (yellow) or Western (orange) diets (total of 340 samples with >800 sequences/sample; see **fig. 1A,B** and **tables S1** and **S5**). The tree was visualized, grouped, and colored using FigTree (http://tree.bio.ed.ac.uk/software/figtree/). Labels: dpc=days post colonization with the human donor sample; dpd=days post diet switch. Sample IDs correspond to those found in **table S1**, unless labeling an otherwise indicated cluster of related samples.

Turnbaugh *et al.*, Supplementary Figure 3

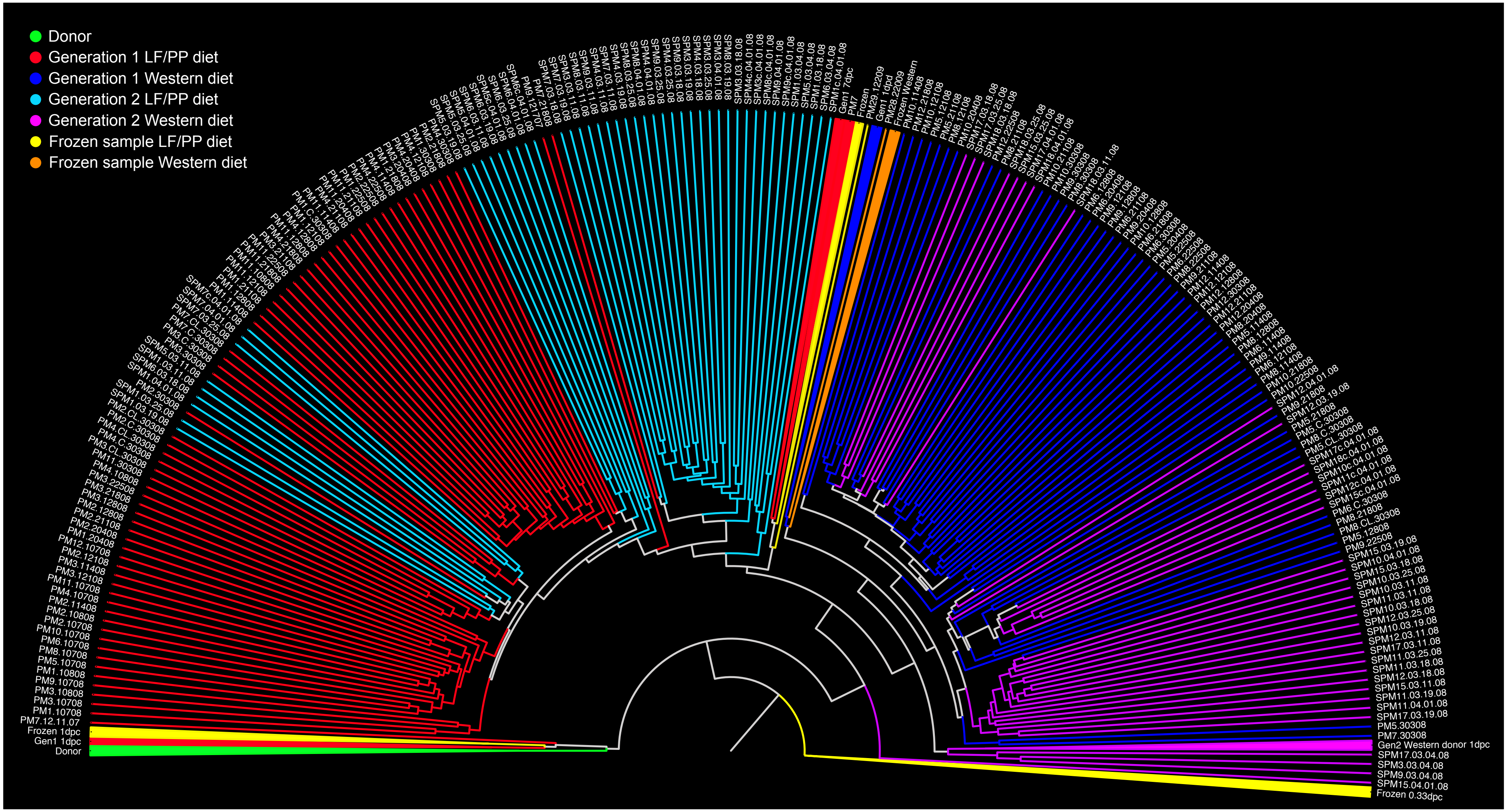


Fig. S4. Assembly of the human gut microbiota in suckling and young adult C57BL/6J mice weaned onto the LF/PP diet. (**A**) UniFrac-based PCoA of V2 16S rRNA gene surveys (weighted analysis; N=146 samples with ≥500 sequences/sample). Percent variance is shown in parentheses. The panel on the left shows time series snapshots of fecal samples obtained from mice between P14 and P24. The middle panel displays PC1 versus PC2, and while in the right hand panel the axes of the PCoA are rotated to show PC2 versus PC3 (blue denotes suckling and young adult mice prior to being individually caged). (**B**) Taxonomic distribution [RDP level 3 (*S18*)] of the gut microbiota sampled from individual mice from postnatal days P14 to P24. (**C**) Evidence for the stability of the fecal microbiota in mice following weaning. Fecal samples were obtained from individual mice every day from P30-P44. UniFrac measurements were performed on various combinations of samples whose collection interval was separated by periods ranging from 0 to 14 days, to measure intra- and inter-individual variation. Pairwise unweighted and weighted UniFrac distances were calculated for all possible comparisons separated by 0 to 14 days. The results indicate that samples from the same mice are no more similar than samples from different mice, and UniFrac distances are generally consistent across the two-week sampling period. Values plotted represent the mean±SEM.

Turnbaugh *et al.*, Supplementary Figure 4

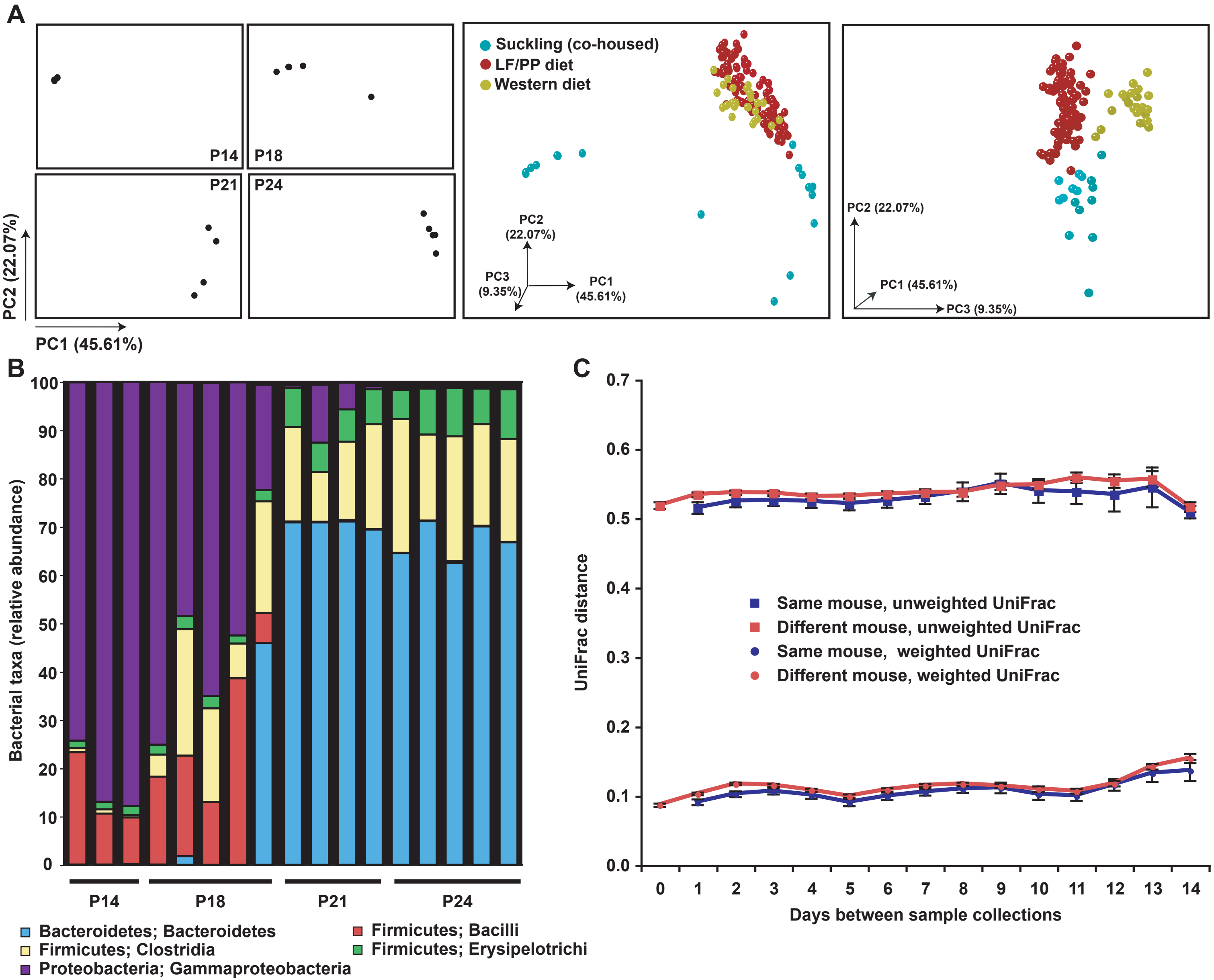
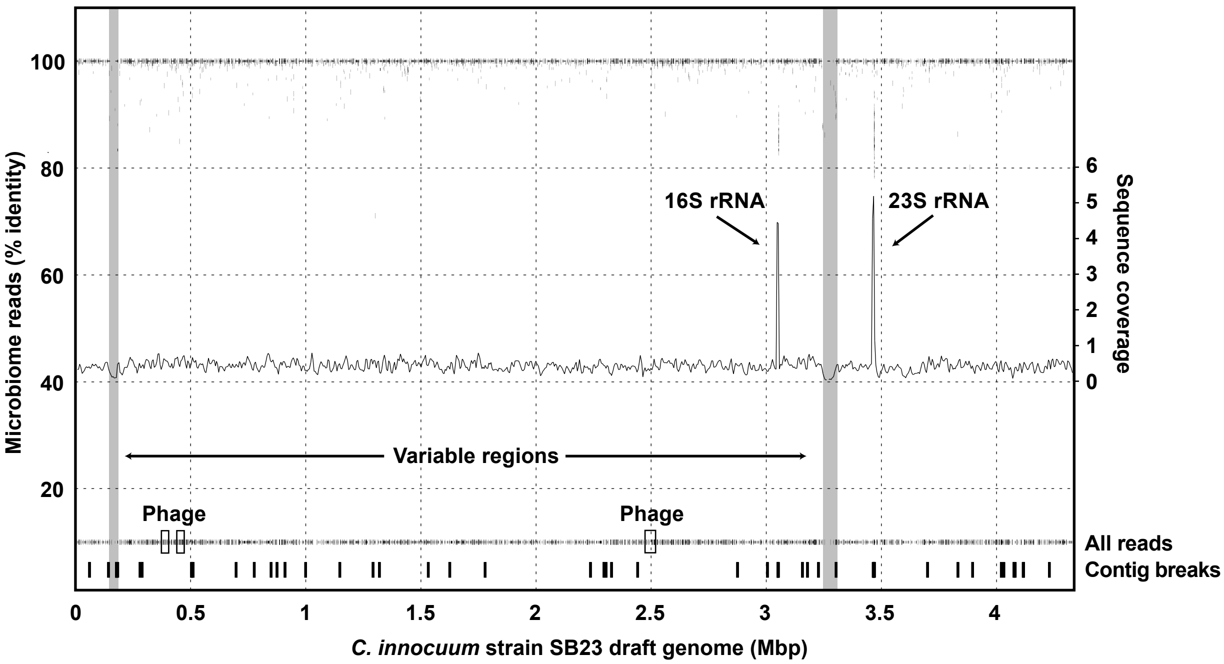


Fig. S5. Analysis of the *in vivo* Erysipelotrichi population, anchored to the *C.innocuum* SB23 draft genome. Annotation of the *C. innocuum* SB23 genome (location on concatenated genomic contigs shown on the x-axis; N=53 contigs). Data on the y-axis represents the percent identity between shotgun fecal microbiome reads (**table S8**) and the strain SB23 genome (dashes near 100%), in addition to the level of coverage (microbiome bases/genome bases across 5kb windows). Additional features, including variable regions, high/low coverage areas, predicted phage, and contig breaks are shown. In general, the level of coverage was relatively even across the reference genome (0.42±0.01), with two contigs having low coverage (representing potential genomic islands), and two regions having increased coverage (containing conserved genes, e.g., encoding 16S and 23S rRNA). The SB23 genome contains three predicted prophage with closest sequence similarity to prophage present in other Firmicutes genomes (including 1 Bacillales and two Clostridia). There are also 6 putative CRISPR regions in the genome, which provide bacterial 'immunity' to phage (*S22*): these CRISPR elements include 14 unique repeat sequences (bacterial machinery), 15 unique spacer sequences (predicted phage elements), and a full genomic region of CRISPR-associated proteins (contig00022, nucleotides 11,826-18,801). Significant matches to both the spacers and repeats were found in the humanized mouse gut microbiome (N=10 and 16 respectively, BLASTN e-value<10^{-5}).

Turnbaugh *et al.*, Supplementary Figure 5



SUPPLEMENTARY TABLE LEGENDS

Table S1. V2 16S rRNA gene sequencing statistics from human donor and two generations of recipient humanized mice.

Table S2. Full-length 16S rRNA gene sequencing statistics.

Table S3. Abundance of class-level bacterial taxa in the gut microbiota.

Table S4. Abundance of genus-level bacterial taxa in the gut microbiota.

Table S5. V2 16S rRNA gene sequencing statistics from transplantation of a frozen human fecal sample.

Table S6. V2 16S rRNA gene sequencing statistics from the assembly of the humanized mouse gut microbiota.

Table S7. V2 16S rRNA gene sequencing statistics from humanized mouse biogeography analysis.

Table S8. Microbiome sequencing statistics (fecal samples).

Table S9. Genomes in the human gut microbe database.

Table S10. Relative abundance of KEGG orthologous groups (KOs) in Western diet associated pathways (% of KO assignments).

Table S11. *Clostridium innocuum* strain SB23 genome sequencing statistics.

Table S12. Number of genes assigned to glycoside hydrolase family 1 in gut Firmicutes and Bacteroidetes.

Table S13. cDNA sequencing statistics.

Table S14. *Clostridium innocuum* strain SB23 genes upregulated in the ceca of humanized mice fed the Western diet relative to the LF/PP diet.

Table S15. Primers used for qRT-PCR analysis.

Table S16. Community gene clusters differentially expressed in the ceca of humanized mice fed the Western diet relative to the LF/PP diet.